SO-101 WAM

An Independent Fine-Tune of Cosmos 3 Edge as a World Action Model: Dreaming, Ordering, and Acting with One Checkpoint

Hugo Hernández
Alakazam · Technical Report · August 2026

On July 20, 2026, NVIDIA released Cosmos 3 Edge [1]: a 4B-parameter, single-GPU world model with a native action pathway, under the OpenMDW-1.1 license. Eleven days later we began fine-tuning it on 254 episodes (101,775 frames; 229 train / 25 held out) of an SO-101 tabletop arm performing one pick-and-place task, through NVIDIA's own cosmos-framework [3], whose core we left deliberately untouched — though no public fine-tuning path existed for a new robot, and the training route itself had to be reconstructed from the framework's internals before the first GPU-hour (§4). Two dated public searches (July 31, August 5) found no other independent fine-tune of this model; NVIDIA's own DROID policy variant exists first-party [2]. This report documents what one fine-tune buys when the same weights are run in every direction the architecture permits, with every claim measured and the failures printed beside the wins. Forward: dreamed episodes decode back to their commanded actions at 20.64 median MAE against a 19.82 floor set by real footage through the same decoder — a pre-registered quality filter that cannot tell dreams from reality, published here as a diagnostic. As a service: another agent ordered 40 episodes at block positions the dataset never visited; 33 were delivered for $1.32 in 72 GPU-minutes, the dreamed block landing a median 1 px from the ordered pixel. Backward: denoising actions from zeros turns the dreamer into a policy that falls 34.77 → 10.84 units over a 17× step range and beats both trivial baselines on the gripper channel (14.92 vs HOLD 28.83, MEAN 18.81) while still failing our pre-registered overall gate against HOLD (10.84 vs 7.05) — a checkpoint on a slope, reported as such. Between: a validated end-effector oracle localizes the model's one persistent defect (the hand at grasp and deposit, wrong already at chunk 0 under perfect conditioning), a controlled sweep exhausts every inference-side fix within 4% of baseline, and a matched-budget experiment refutes the wrist-camera retrain everyone reaches for (+9% end-effector error at equal tokens and steps). On hardware, five days in: the policy commanded its first grasp on a real camera frame (gripper 39.3 < 50) after 375 consecutive render-conditioned chunks in which it never once did — the render domain gap measured from both sides. Total metered spend: ≈$342 of a $500 pre-registered ceiling, spot instances only.

Release: `alakazam.gg` (landing, demo clips, blog) · **Base model:** `huggingface.co/nvidia/Cosmos3-Edge`

Report sources: a campaign ledger of 51 numbered iterations; every number below traces to a logged artifact. Projections are marked as estimates.

Contents

1 Introduction and claims

A world action model (WAM) denoises video and action trajectories jointly: conditioned on frames and actions it predicts futures; conditioned on frames alone it can write the actions. Cosmos 3 Edge ships this pathway natively. The question this campaign set out to answer is narrow and practical: *what does one fine-tune of a 4B edge-class WAM actually buy, on one cheap robot, measured honestly?* Not "can it produce a demo reel" — what does it do under instruments, at which budget, and where exactly does it stop.

The method underneath every section is the same: pre-register the gate, build the instrument before trusting the number, keep the boring baseline in the table, and let the ledger — 51 numbered iterations with receipts — be the source of truth. Where a result is two points on a curve we say "two points"; where a claim rests on absence of evidence we date the searches. Table 1 states every headline claim with its evidence class.

#	Claim	Evidence class
C1	First independent fine-tune of Cosmos 3 Edge	Absence of contrary evidence: two dated public searches (Jul 31, Aug 5, 2026); NVIDIA's DROID policy variant is first-party [2]
C2	Dreams obey commanded actions	Measured: cycle-decode median 20.64 (n=39) vs 19.82 real floor (n=12), same decoder, same chunk indices (§5.3)
C3	The residual defect is in-model at chunk 0; inference-side fixes are exhausted	Measured: controlled A/Bs, 3 hand-offs × guidance {1,2,3,5}, all within 4% of baseline; chunk-0 EE error among the episode's worst (§6)
C4	A wrist camera does not pay at matched budget	Measured: identical token count and steps, +9% end-effector, +40% global vs front-only (§7)
C5	The policy beats trivial baselines on the gripper channel and fails our overall gate against HOLD	Measured: 14.92 vs HOLD 28.83 / MEAN 18.81 (gripper); 10.84 vs HOLD 7.05 (overall) = registered FAIL; curve not flattened at budget stop (§8)
C6	The MuJoCo rung is fair on physics and closed by rendering	Measured: human demos 14–17/20 under a pre-registered bar; policy 0/5 on corrected renders; gripper never <50 in 375 render chunks (§10)
C7	The policy is alive on real imagery	Measured: first real-frame inference commanded gripper 39.3 < 50 with the arm sweeping toward the block, under ±250-tick clamps (§11)
C8	The fine-tune is a usable synthetic-data service	Measured: 33/40 episodes delivered to another agent's spec; $1.32; 1 px median anchor adherence; 1.0 cm median image/action agreement (§9)

Table 1 — the release's claims, each tied to the section that measures it. Nothing here rests on a loss curve or an eyeball.

2 The model and its action pathway

Cosmos 3 Edge is a dual-tower mixture-of-transformers: a reasoning tower that encodes the visual context and a generation tower that denoises outputs, ~4B parameters total, BF16, designed for single-GPU operation [1]. The action pathway attaches action tokens to the same denoising process. Three operating modes matter for this report, and they differ only in conditioning masks:

- **Forward dynamics (FD)** — condition on frames + *all* actions, denoise video. In the framework, FD sets `condition_frame_indexes_action = list(range(action_length))` (transforms.py:321) — every action token is conditioning, so *the action head receives no gradient in pure FD training*. This detail decides §8's design.
- **Inverse dynamics (ID)** — condition on frames, denoise the actions that connect them. Used as the decoder in §5.3.
- **WAM / joint** — denoise video and actions together; the framework's `mode="joint"` randomly assigns FD/ID/WAM per sample (cosmos3_action_lerobot.py:744), so action supervision arrives on roughly two-thirds of samples while forward dynamics is retained.

One architectural fact closed an entire avenue early: the classifier-free-guidance unconditional branch differs from the conditional branch *only in the text tokens* (omni_mot_model.py:2631–3036). Guidance can amplify prompt adherence; it cannot amplify action adherence, because actions are identical on both branches. Any hope that a guidance knob would fix an action-following defect was dead on inspection, and §6 confirms it empirically.

3 Data

The corpus is one task, one scene family: an SO-101 [5] picks a blue block from a mat and deposits it in a yellow bin, recorded at 640×480 / 30 fps with a fixed front camera, actions stored in LeRobot format [4] — six channels (five joints + gripper), normalized to [−100, 100] per joint with the gripper effectively in [0, 100] (≈2 closed, ≈68–70 open). We write "units" throughout; the corpus spans ≈200 units on arm channels. The trainer consumes these values raw (`action_normalization=None`, the vendor's FD default).

Composition. 174 real episodes, plus material-restyled and distractor-augmented variants produced by our Forge pipeline, totalling **254 episodes / 101,775 frames** after leak filtering: any augmented episode whose source lay in the held-out split was dropped (restyle sources 52, 143; distractor sources 24, 45, 108, 164). Split: **229 train / 25 held out**, the held-out set fixed once and reused by every gate in this report.

Alignment defects found and fixed before training (each would have silently degraded action conditioning):

- Concatenated augmentation videos carried `avg_frame_rate = 29.99997`; over an episode this drifts one full frame against the 30 fps action track. Re-encoded CFR (`-vsync cfr -r 30`) rather than widening the pairing tolerance — a tolerance band-aid would have hidden the same class of bug forever.
- macOS AppleDouble `._*` files shipped inside a tarball were globbed as parquet inputs (UnicodeDecodeError); deleted, and `COPYFILE_DISABLE=1` adopted.

- A `tasks.parquet` written via `pyarrow.table()` lacked the pandas index metadata the loader expects; the task string resolved to the integer 0. Isolated with a class×dataset cross-test, fixed by writing through pandas with an explicit index.

4 Training runs

The path that did not exist. Cosmos 3 Edge shipped as weights and an inference route. The framework carries post-training recipes for NVIDIA's own robots; for a *new* robot — new action space, new embodiment — there was no public path to execute, and more than one mechanism we expected from documentation and early reconnaissance turned out not to exist in the shipped code. Before the first GPU-hour, the fine-tuning route had to be reconstructed from the framework's internals: embodiment registration, the dataset interface, the experiment configuration, and a verification harness around them. Three separate times during that reconstruction, something that looked functional was a silent no-op — a run that would have looked healthy while training nothing — and was caught only by an executable assertion. We are deliberately light on the specifics; the point of record is that this reconstruction, not GPU time, was the campaign's first hurdle, and it is a fair share of what "first independent fine-tune" means in practice.

All training ran through the framework's registered-experiment system — its core unmodified — on single spot GPUs (Table 2). Sequence packing is the load-bearing lever: the packer fills a fixed token budget (`max_num_tokens_after_packing = 74000`), so sample cost scales with pixel count. Dropping from the 480 bucket (736×544 = 400,384 px) to the 256 bucket (320×256 = 81,920 px) packs ≈4.9× more clips per step — the correct trade whenever the target is actions rather than pixels, and the reason the policy arms trained in hours, not days.

Run	Objective	Steps	Bucket	Warm start	Box
Video arm (ft7600)	FD	7,600	480	base weights	A100-40 spot
Concat probe	FD, two-view	900	480 (2×half-res)	base weights	RTX Pro 6000 spot
Concat extension	FD, two-view	+6,700	480 (2×half-res)	concat@900	RTX Pro 6000 spot
Policy probe	pure WAM	450	256	ft7600	RTX Pro 6000 spot
Policy stage 1	joint	1,900	256	probe@450	RTX Pro 6000 spot
Policy stage 2 (s2)	joint	+5,700	256	stage1@1900	RTX Pro 6000 spot

Table 2 — training runs. The RTX Pro 6000 ran the identical config 1.85× faster than the A100-40 (3.5 vs 6.5 s/it). "+N" = continuation to a 7,600-step-equivalent budget.

Warm-start mechanics that cost us a day to learn. The framework treats a checkpoint directory as a resumable run: pointing `load_path` at it restores the optimizer and — worse — the LR schedule, so a "fresh" run resumes at a dead learning rate. The fix is a weights-only view: symlink just the `model/` subtree into a fresh directory and start a new schedule with `cycle_lengths = [max_iter]` (any mismatch raises a KeyValidationError). Every continuation in Table 2 used this. The concat probe taught the corollary: its `cycle_lengths=[900]` meant the 600→900 improvement we initially reported was measured *on a dying LR*, understating the arm's true slope — one reason the matched-coverage extension (§7) was mandatory before any verdict.

Operating rules baked into the launchers, after each bit once: per-save GCS sync (a preempted box must never hold the only copy); warm-start source directories self-delete once training passes iteration 5 (two disk-full incidents came from counting only the active run's checkpoints); `HF_HUB_OFFLINE=1` after the first weight pull; framework jobs never run as root.

5 Instruments

Every verdict in this report is produced by an instrument that passed its own validation gate first. Five did the work; three earlier candidates were killed by their gates and discarded. This section is the report's core, because the instruments are the reusable part.

5.1 The end-effector oracle

No published work in our citation set isolates end-effector orientation as an ablated metric (verified against [6–10]), so we built one. Background subtraction (per-pixel median over the episode = the static scene) isolates the foreground; a flood fill keeps the component connected to the arm's fixed mount; the tip is the component's farthest extent. No camera calibration required — we have none.

Validation before use: detected tip-y against *commanded* joints across the corpus: r = +0.858 (wrist_flex), −0.827 (elbow_flex), +0.787 (shoulder_lift); re-fit at native scale r = +0.817. The instrument provably tracks the real arm. Two predecessors failed exactly here: a color detector resolved to the blue-tinted mat (169k px of table), and a brightness detector inside a rectangular ROI pinned its tip to a corner (r = 0.131). A third — a block tracker built later for the anchor-leak test — failed *its* ground-truth check and was discarded before it could vote. The gate is the method.

Metric: mean $|\text{dreamed} - \text{real}|$ in a window centred on the *real* end effector (0–255 scale). EE-local error runs ≈2× the global frame error: the hand is the worst-rendered region, which is why global MAE never saw the defect.

5.2 Alignment-safe scoring

Each generated chunk reproduces its conditioning frame as frame 0; naive concatenation yields 255 frames for 240 actions and silently shifts every comparison one frame per chunk. All scoring drops frame 0 per chunk (chunk c compares frames 1–16 against real frames 16c+1…16c+16), and all cross-sequence comparisons are cropped from a single composed video so scaling is byte-identical across arms. An early mislabelled 4-up (a name→label map bug tagging the motion arm "BASELINE") was caught before publication; a mislabelled comparison is worse than none.

5.3 The cycle-consistency floor

To ask "did the dream follow its orders?" without a human in the loop, we decode actions back out of dreamed video with the joint checkpoint in ID mode and compare against the commanded trajectory. The critical control is the **floor**: real footage pushed through the same decoder at the same chunk indices scores 19.82 median MAE (n=12) — that is decoder error, not dream error. Dreams score 20.64 (n=39), p90 23.35 vs 21.91, per-channel structure near-identical, separability AUC 0.596. A rule registered *before* the floor was known said: if the filter cannot discriminate dreamed from real, it culls nothing and ships as a diagnostic. It fired. We publish the number as obedience evidence (C2) and refuse it as a filter.

5.4 World-to-pixel calibration without a camera model

The simulator's fitted camera (§10), projected directly onto real 480p frames, misses by 305 px (84 px after a best-fit 2-D similarity) — a fitted render camera is not a photogrammetric solution. For the data service we instead fitted a homography on 120 real correspondences (simulator forward-kinematics pinch point at the real grasp instant ↔ detected block pixel): leave-one-out 6.3 px, p90 8.4 px. The quantity that matters downstream — differential displacement over ≤8.5 cm moves — agrees at **7.2 px ≈ 1.0 cm median** (12.9 px ≈ 1.9 cm p90, 332 pairs). Every §9 claim cites this number, not the raw fit.

5.5 Boring baselines

Every policy number is printed beside HOLD (repeat the last observed action) and MEAN (the corpus average). §8.1 shows why this is non-negotiable: a loss curve fell 33,590 → 12,000 and looked excellent while the policy was losing to *both* baselines on every arm channel. On a 1.1-second horizon, standing still scores 7.05. Any action paper that omits HOLD is reporting against air.

6 Dream fidelity: where the hand is wrong

The video arm's dreams track the task through reach, transit, and deposit from a single conditioning frame plus 240 actions. The oracle localizes what remains wrong, chunk by chunk (Figure 1): error is worst at **grasp (chunk 0: 21.9)** and **deposit (chunks 13–14: 27–35)** and best in transit — the "gross motion right, fine manipulation wrong" signature the survey literature documents across the field [6].

Figure 1 — end-effector error per chunk on the baseline rollout, held-out ep52. Chunk 0 has a real photograph as conditioning and zero accumulated drift, and is still among the worst chunks in the episode.

Chunk 0 is the decisive measurement. It is identical across all hand-off variants (21.88/21.88/21.91 — as it must be, since all condition on the same real frame), it contains zero drift, and it is nearly the worst chunk in the episode. **The hand is wrong before any chaining begins.** No rollout-side change can repair error produced inside the first chunk.

We confirmed that exhaustively rather than rhetorically (Figure 2, Table 3): three chunk-to-chunk hand-off strategies (baseline one-frame×16; anchor re-injection, imported from our own camera-navigation serving stack; motion-continuous, feeding the last five generated frames in order) and classifier-free guidance at {2, 3, 5} on the best hand-off. Everything lands within 4% of baseline. The anchor variant is 23% *worse*, and the mechanism is measurable: the anchored dream's second half sits *closer to the opening frame than reality does* (8.74 vs real 11.04; baseline 14.01) — a sixteen-second-stale anchor is contradictory evidence for a scene that must change, and it drags the moved block back toward its start pose. A technique that stabilized a static world under a moving camera destabilizes a changing world under a fixed one. Verdicts do not cross task classes.

Rollout recipe	EE-local MAE	vs baseline
guidance 3.0 + motion hand-off	19.51	−3.7%
guidance 5.0 + motion hand-off	19.82	−2.2%
motion hand-off	19.85	−2.0%
guidance 2.0 + motion hand-off	19.93	−1.7%
baseline (one frame ×16)	20.26	0.0%
anchor re-injection	24.96	+23.2%

Table 3 — the inference-side book, exhausted. Same checkpoint (ft7600), same actions, same seed, held-out ep52, one instrument at native scale. §2 predicted the guidance rows: the unconditional branch differs only in text tokens.

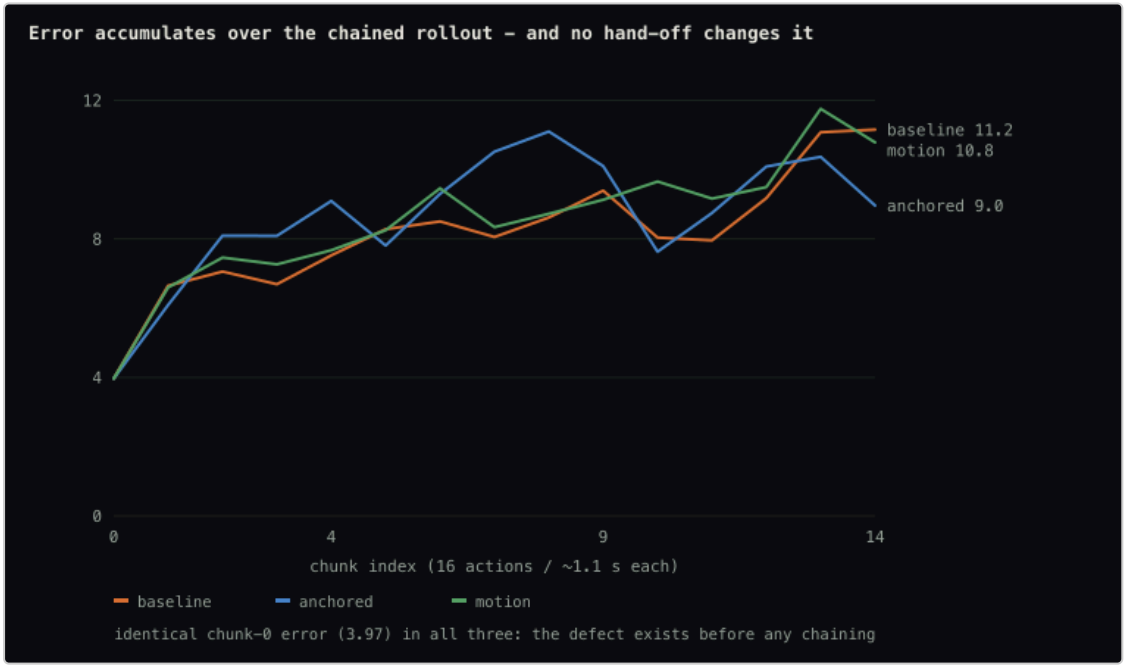

Figure 2 — global error compounds ≈3× over fifteen chained chunks (3.97 → 11.2) and every hand-off rides the same curve: the error is manufactured inside each chunk, not between chunks. Identical chunk-0 values pin the defect upstream of chaining.

> **Measurement order matters.** Global MAE initially ranked the three hand-offs "identical" and the motion hand-off *worse* than baseline (8.515 vs 8.143). The EE-local oracle reversed that: motion is the best recipe at the end effector. A metric averaged over the whole frame outvoted the 4% of pixels where the task lives. We re-scored everything through the localized instrument before any verdict survived.

7 The wrist camera at matched budget

With inference exhausted, the fix must be training-side, and the field's default prescription is an eye-in-hand camera. Our own measurement supports the mechanism: per unit of wrist motion, the wrist view moves 5.7× more pixels than the front view (sensitivity 0.810 vs 0.141), and in the wrist view a roll error moves *every* pixel. The literature is directionally unanimous [7, 8]. The question a budget-bound model must ask is different: **at fixed tokens, does splitting the pixel budget across two views beat spending it all on one?**

The design keeps token cost exactly neutral: stack the two views vertically into the (544, 736) bucket — the identical pixel count to the front-only (736, 544) layout — each view at 544×368. Train to the same 7,600 steps. Score with the same oracle on the same held-out episode.

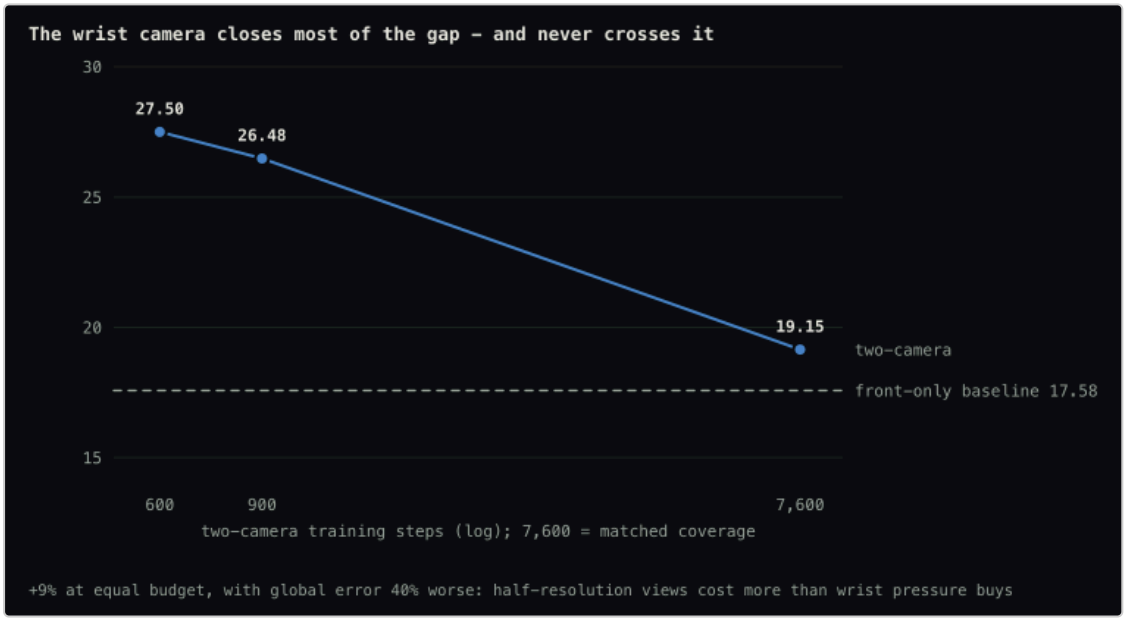

Figure 3 — the two-camera arm closes most of its early deficit (27.50 → 19.15) and finishes 9% behind the front-only baseline (17.58) it was meant to beat, at identical token count and steps. Global error is 40% worse: the halved front resolution costs everywhere.

Two process notes that made this result trustworthy. First, the 900-step probe was *not* allowed to stand as a verdict — at 900 steps the arm was still learning to draw a new output geometry (the wrist half rendered as visible smear), the improvement slope was measured on a dying LR (§4), and we wrote "UNTESTED, two points" in the ledger rather than "refuted". The refutation only exists because the extension ran to matched coverage. Second, the two views have different white balance (the bin reads orange in the wrist view, yellow in front); any color-based metric would need per-half calibration — one more reason the oracle is geometry-based.

The reconciliation with the literature is a budget statement, not a contradiction: published wrist-camera wins add a view *and add tokens*. Under a fixed budget — which is what "edge" means — the trade inverts, at least at this scale and task. This is the campaign's most reusable negative result.

8 The policy program

8.1 The probe, printed in full

The first attempt ran 450 steps of pure-WAM training from the video checkpoint and failed the gate on every channel. We print it in full because its shape drove every design decision after it — and because a report that only shows the fixed version is advertising.

Channel	Policy@450	HOLD	MEAN
shoulder_pan	27.16	1.08	27.76
shoulder_lift	41.34	5.10	46.64
elbow_flex	37.69	2.70	42.90
wrist_flex	42.84	3.10	15.02
wrist_roll	39.72	1.50	32.10
gripper	**19.88**	28.83	18.81
overall	**34.77**	**7.05**	**30.54**

Table 4 — the 450-step probe: 25 held-out grasp-centred windows, actions denoised from zeros. FAIL against both baselines on every arm channel — while the training loss fell 33,590 → ≈12,000 and looked excellent. One channel carried signal: the gripper already beat HOLD, on the only channel where holding is a poor predictor.

Diagnosis: predicted arm channels clustered near zero (spans of 2–5 units against ground truth spanning tens) — a rectified-flow head, denoising from N(0,1) toward raw-scale targets, had not yet learned a ≈50× amplification. Two mechanical facts made the probe non-fatal: byte-compare at iteration 101 proved the action parameters really train, and the pure-FD gradient analysis (§2) says a video-arm checkpoint contains an *untouched* action head — 450 steps was always a probe, not an attempt.

8.2 Staged budget with registered decision rules

The real attempt changed three things, each with a written reason: `mode="joint"` (action supervision on ~2/3 of samples while retaining the certified forward dynamics — one checkpoint must keep dreaming); warm start from the probe (verified by step-1 loss: 14,684, the probe's ending level, vs 33,590 fresh); and a staged budget with rules registered before results — stage 1 (1,900 steps) must cross MEAN to unlock stage 2 (to 7,600-equivalent), else the pivot was normalization, not more steps.

Gate	Overall	Gripper	Rule outcome
probe @450 (25 ep)	34.77	19.88	FAIL; probe only
stage 1 @950 (10 ep)	25.60	18.9	crossed MEAN → continue
stage 1 @1900 (25 ep)	20.18	19.9	rule fired → stage 2
stage 2 @7600-eq (25 ep)	**10.84**	**14.92**	overall gate vs HOLD: **FAIL**; gripper beats both baselines

Table 5 — the staged program against the same 25-episode instrument (10-ep mini-gate at 950 marked as such). Per-channel at 1900 vs the probe: pan 27.2→9.7 (−64%), wrist_flex 42.8→16.4 (−62%). The gripper sat flat (19.9/18.9/19.9) through stage 1 and broke loose only in stage 2 (14.92).

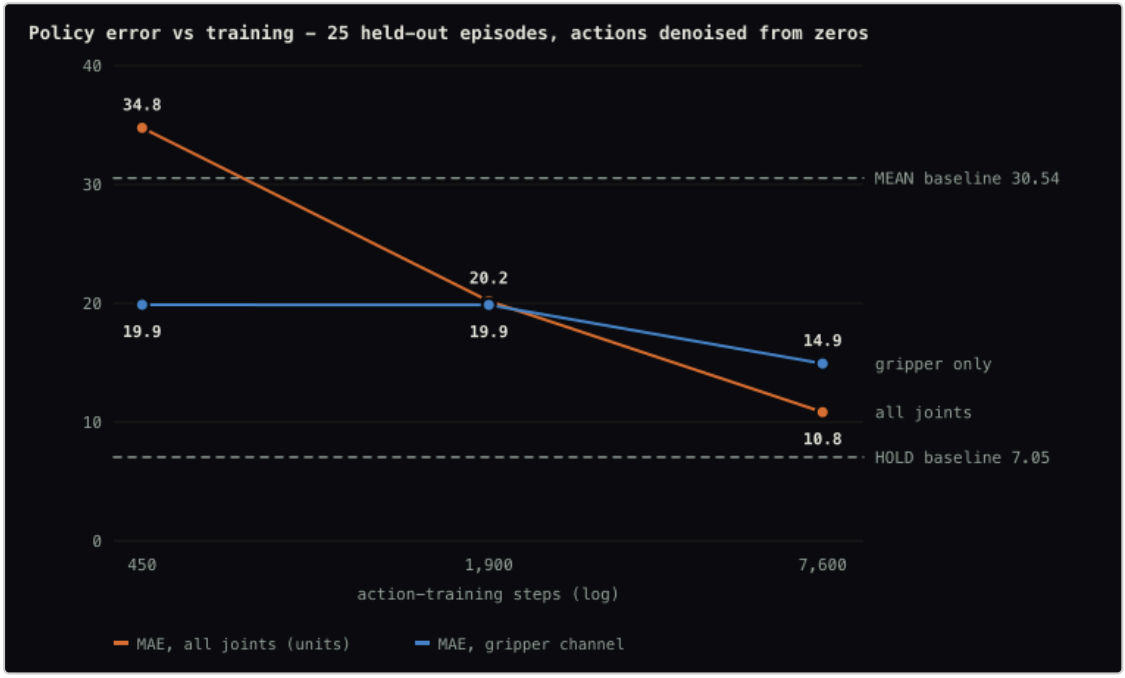

Figure 4 — the headline curve. Overall error falls 3.2× across a 17× step range and has not flattened at budget exhaustion. Dashed lines are all-joints baselines; the gripper-channel baselines are HOLD 28.83 / MEAN 18.81. We report a checkpoint on a slope.

The honest reading, in both directions. Against the pre-registered overall gate the policy **fails**: 10.84 vs HOLD's 7.05, because slow arm joints reward standing still on a 1.1 s horizon. On the gripper — the channel that decides whether a grasp happens, and the one where HOLD is structurally poor — it beats both baselines by a clear margin. The gripper's own trajectory (flat through 1,900 steps, then −25% in stage 2) had been flagged at the 1900 gate as the channel to watch, with conditioning history (`max_num_history_actions`) pre-registered as the pivot if it stayed flat. It did not stay flat.

Deployment shaping. The head's raw gripper output hedges between the corpus's two modes (open ≈68–70, closed ≈2) rather than committing — the signature of a rectified-flow head averaging a bimodal target. At deploy time we binarize at 50 (≤50 → full close). This is a decision rule applied to a measured distribution, not a training fix; §11 shows it acting on hardware.

8.3 The closed imagination loop, published as a failure

Policy proposes a chunk, dreamer renders it, policy reads the dream and proposes again — training-in-imagination is the obvious end state for a WAM, so we ran it and publish the outcome: it does not work at this scale today. Two failure modes compound. The 256-bucket dreams degrade over iterations (the policy arm's resolution was bought for action throughput, §4, and the loop inherits the bill), and the gripper hedge (§8.2) becomes a fixed point inside a blurring dream: over 15 rounds on model-rendered frames the commanded gripper plateaus at 55.7–56.8, never crossing the close threshold, while the same checkpoint on *real* frames dives to 36–39 within two chunks. There is nothing firm in the dream to binarize against. The loop needs either a higher-resolution dreaming arm or a commitment mechanism in the head; we name both and claim neither.

9 The synthetic-data service

Midway through the campaign, a separate agent (another Claude session running a navigation-policy program) asked whether our checkpoint could generate pick-and-place episodes at block positions its corpus lacked. We formalized the interface — a written generation brief with the units contract, the capability envelope, and yield-line reporting — and filled the first order. The pipeline: composite the block at the requested pixel (median-fill the hole), harmonize the composite with an image model (30/40 passed a geometry gate after one retry; 9 fell back to the raw composite; 1 failed outright and was dropped), then chained FD generation, then pre-registered filters.

Stage	Count	Note
ordered	40	anchor positions from the requester's spec
anchors passing geometry gate	39	30 harmonized + 9 raw-composite fallbacks
generated	39	240 frames = 240 actions each (frame-0 rule, §5.2)
passing EE filters	**33**	filters calibrated on real episodes first: 21/24 pass

Table 6 — the delivery. Filters (block trackable ≥0.70, displacement ≥120 px, terminal position ≤200 px from bin) were pre-registered against real episodes before touching synthetic ones. Cost: 72 minutes of one spot GPU, $1.32 uptime-derived.

The two numbers a downstream trainer should care about: the dreamed block sits a **median 1 px** from the ordered pixel in frame 0 (the model honors a painted anchor), and delivered image/action agreement is **1.0 cm median** through the §5.4 homography. The cycle filter (§5.3) ran as a diagnostic per its registered rule and culled nothing.

Filling the order also audited the requester's own corpus, for free: in 80/80 of their retargeted MimicGen shards, the IK-retarget boundary lands *after* the deposit event (median +85 frames past it), because a pinch-height heuristic had selected the post-deposit retraction as the grasp instant — offsetting entire trajectories against a bin that never moves. Our fill takes the release instant from the real gripper trace instead; carry and deposit are byte-identical to the source demo in 40/40, and yield doubled against their pipeline (40% vs 15–20%). A generation service that can read its requester's data earns its keep twice.

10 The simulation gate

Before hardware we built the cheap rehearsal in MuJoCo (scene and arm from the public SO-ARM100 menagerie assets, our camera fitted to the real rig: az 50°, el −84°, dist 0.7 m, fovy 58°, furniture rotated to match). The protocol is self-consistent placement: roll the policy once with a parked probe, anchor the trial at the close-after-open gripper event, place the object at the pinch point with a 3×3 micro-grid (±1.5 cm), and score lift > 3 cm followed by carry > 5 cm. The bar was written before any run.

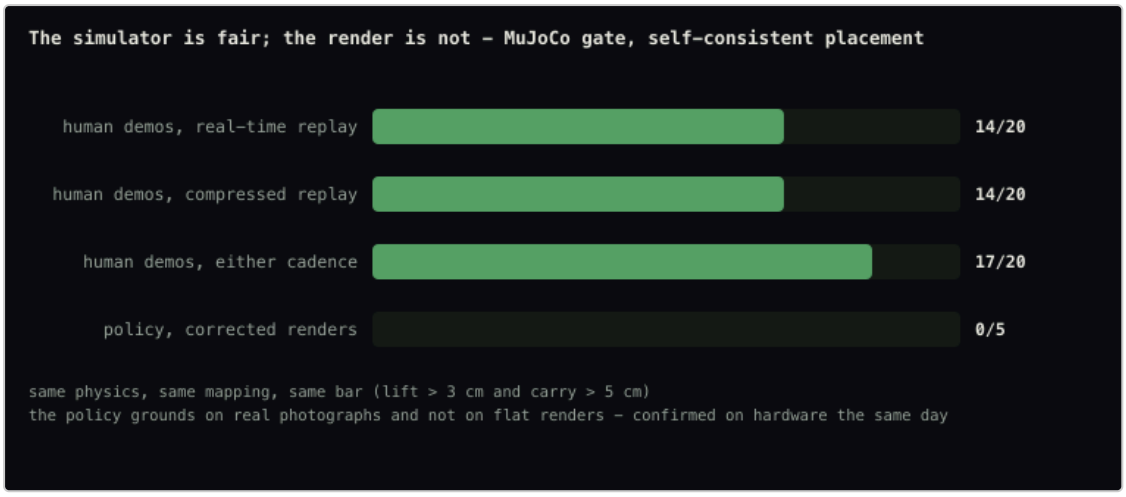

Figure 5 — the gate. Replayed human demonstrations pass 14/20 in real time, 14/20 compressed, 17/20 under either cadence: physics, mapping, and bar are all fair. The policy on corrected renders: 0/5.

The failure is specific and diagnostic: on rendered frames the policy's gripper channel never left its open plateau (≥56 units) across 375 consecutive chunks — zero grasp attempts, while §11 shows the same checkpoint diving below the close threshold on its first real photograph. The renders are geometrically corrected but texturally flat, and the rendered arm is bare white plastic where the real arm is taped and wired — the largest object in view, maximally wrong. We close the sim rung with a verdict about rendering, not physics: the harness (lockstep batching, 25-way, zero drops) is ready the day a render-robust policy exists.

> **A stale file almost shipped a wrong verdict.** The first policy-in-sim eval scored 0/25 against a camera pose read from a snapshot file that a peer session had already retracted (az 285°/el −8°). The corrected pose came from the live scene. Rule banked: regenerate derived artifacts from the live source; a plausible cached file is the most expensive kind of wrong.

11 Hardware

Day five. The rig: an SO-101 on a Jetson (Feetech STS3215 bus servos, 1 Mbaud), a fixed camera served as MJPEG, the s2 checkpoint on a cloud GPU, and a laptop bridging the two over ssh — the Jetson and the GPU cannot reach each other, so the loop runs at ≈90 s per decision round. Safety by construction, not by hope: every joint goal clamped to ±250 ticks from the current pose, a ±900-tick fence around the home pose, gentle acceleration/speed registers, torque released whenever the arm is left unattended. Because the recording rig's calibration file is not in our possession, absolute joint targets are undefined; the bridge therefore applies *intra-chunk motion deltas* — the unknown offset cancels, and scale error is absorbed by a bounded gain.

One trap worth printing: on arrival, torque read *enabled* on all six servos despite a peer session's note saying released. A goal write would have moved a live arm. Re-read the state you were promised.

The result that matters (C7): on the first real camera frame, the policy commanded gripper **39.3** — below the binarization threshold, a grasp attempt — with elbow span 62 and wrist_flex span 42, and the arm moved under its clamps (j2 −223, j3 +252, j4 +248, j5 −247, j6 +247 ticks). The same checkpoint never commanded below 56 in 375 render-conditioned chunks (§10). The domain-gap verdict is now confirmed from

both sides on the same day: alive on photographs, blind on renders. No completed pick yet: the camera is mounted side-level where training data looks steeply down, and mid-run the MJPEG feed was found frozen (one stalled JPEG serving repeatedly — telemetry, not the before/after photos, proved the arm had moved). Camera remount and the calibration file are the remaining blockers, and both are logistics.

12 Cost ledger and operating rules

The campaign ran under a pre-registered envelope ($500 hard / $400 planned) with a standing rule that no GPU starts without a priced ledger entry. Meter: ≈**$342** at close, uptime-derived where the billing export lags (it lags 6–24 h; the dashboard is truth at settlement). Documented increments:

Lane	Cost	Note
video arm ft7600 + bring-up + gate generations	≈$180–225	A100-40 spot, the campaign's largest line
hand-off / guidance A-Bs	<$3	RTX Pro 6000 spot, ≈45 min
concat probe + extension to 7,600	≈$36	$8 probe + $28 extension
policy probe + stage 1 + stage 2	≈$34	$2 + $9 + $23
synthetic-data delivery	$1.32	72 min, one spot GPU
simulation + hardware bridge inference	≈$6	CPU sim box + wam-1 restarts
total (meter)	**≈$342**	vs $400 planned / $500 hard

Table 7 — spend by lane. Individual lanes are ledger-stated increments; the total is the metered envelope figure, which includes restarts, smokes, and idle-guard margins not itemized above.

Rules that earned permanence: short verdict jobs run on-demand, not spot (a preemption with `--instance-termination-action DELETE` destroyed a sim box and its disk mid-verdict — on-demand CPU for an hour costs $0.09); every box has a named stop actor before it starts; disk budgets count every checkpoint on the box, not just the active run's; and no result enters the ledger from a container-list snapshot — meters only.

13 Related work

First-party. Cosmos 3 Edge and its DROID policy variant [1, 2] establish the intended post-training path; our contribution is the first documented traversal of it from outside, plus the instrument stack the vendor does not ship. The cosmos-framework [3] was used unmodified — deliberately, so that every finding transfers to the next user of the same code.

Multi-view world models. Ctrl-World [7] jointly predicts multiple views (token-dimension concatenation, zero-shot to novel placements) and its own same-row ablation prices the wrist view at 19.18→15.94 PSNR; WristWorld [8] generates wrist views from anchors. Both add tokens with the added view; §7 measures the fixed-budget version of the question and gets the opposite sign. DualDiff [9] and the event-aware EA-WM [10] weight foreground/manipulation regions in the loss — attractive for the §6 defect — but both require camera calibration we do not have; they are named as the training-side candidates the oracle would gate.

Failure taxonomy. The survey [6] documents "gross motion right, fine manipulation wrong" as the recurring defect class in robot world models; §6 reproduces it under a validated local instrument and adds the chunk-0 localization that separates it from drift.

14 Limitations

- **One task, one scene family.** 254 episodes of a single pick-and-place; no generalization claim of any kind is made or implied. Foreign-data evaluation is the standing requirement for one.
- **MAE is not success rate.** Every policy number is open-loop trajectory error on held-out windows; the only closed-loop successes reported are replayed demonstrations in sim. No completed autonomous pick exists yet on hardware (C7 is a commanded grasp, and we word it that way).
- **The overall policy gate FAILS.** 10.84 vs HOLD 7.05. We publish the gripper win inside that failure, not instead of it.
- **Evidence sizes are small where they are small:** 25 held-out windows, one held-out episode for rollout A/Bs (chosen blind, but n=1), 5-episode render eval, n=39/12 for the cycle floor. Single seeds throughout; no variance bars.
- **The imagination loop does not close** (§8.3), and the wrist defect is located but not fixed — the two named training-side candidates [9, 10] are blocked on calibration we lack.
- **Delta control on hardware** is a workaround for a missing calibration file, and the camera pose does not yet match training. The hardware result is an aliveness proof, not a deployment.
- **"First independent"** is an absence-of-evidence claim with two dated searches behind it, and it expires the day someone else publishes.

15 Conclusion

One fine-tune of a 4B edge world model, priced at a used-laptop budget, is simultaneously: a dreamer whose episodes obey their orders to within the decoder's own floor; a data factory that filled another agent's order at $0.04 per accepted episode; and, run backward, a policy whose gripper channel already beats the baselines that expose most action papers — carried to a first commanded grasp on a real arm five days after training began. The same instruments that establish those wins locate the edges without flattery: a hand drawn wrong at the moment of contact, an imagination loop that cannot yet feed itself, a render gap the policy will not cross, and an overall gate it does not pass. We think the instrument stack — the oracle, the floor, the matched-budget protocol, the boring baselines, the registered rules — is the transferable artifact. The checkpoint is what it measured: the first independent Cosmos 3 Edge, and a map of exactly how far that title gets you.

References

1. NVIDIA. *Cosmos3-Edge* (model card). Hugging Face, July 20, 2026. `huggingface.co/nvidia/Cosmos3-Edge`. OpenMDW-1.1.

2. NVIDIA. *Cosmos3-Edge-Policy-DROID* (model card). Hugging Face, 2026. huggingface.co/nvidia/Cosmos3-Edge-Policy-DROID.
3. NVIDIA. *cosmos-framework* (post-training monorepo). github.com/NVIDIA/cosmos-framework.
4. Hugging Face. *LeRobot*. github.com/huggingface/lerobot.
5. TheRobotStudio & Hugging Face. *SO-ARM100* / *SO-101*. github.com/TheRobotStudio/SO-ARM100.
6. Survey: world models for robotic manipulation. arXiv:2606.00113, 2026.
7. Ctrl-World: controllable multi-view world models. arXiv:2510.10125, 2025.
8. WristWorld: wrist-view generation for manipulation. arXiv:2510.07313, 2025.
9. DualDiff: foreground-weighted diffusion for driving world models. arXiv:2505.01857, 2025.
10. EA-WM: event-aware world modeling. arXiv:2605.06192, 2026.